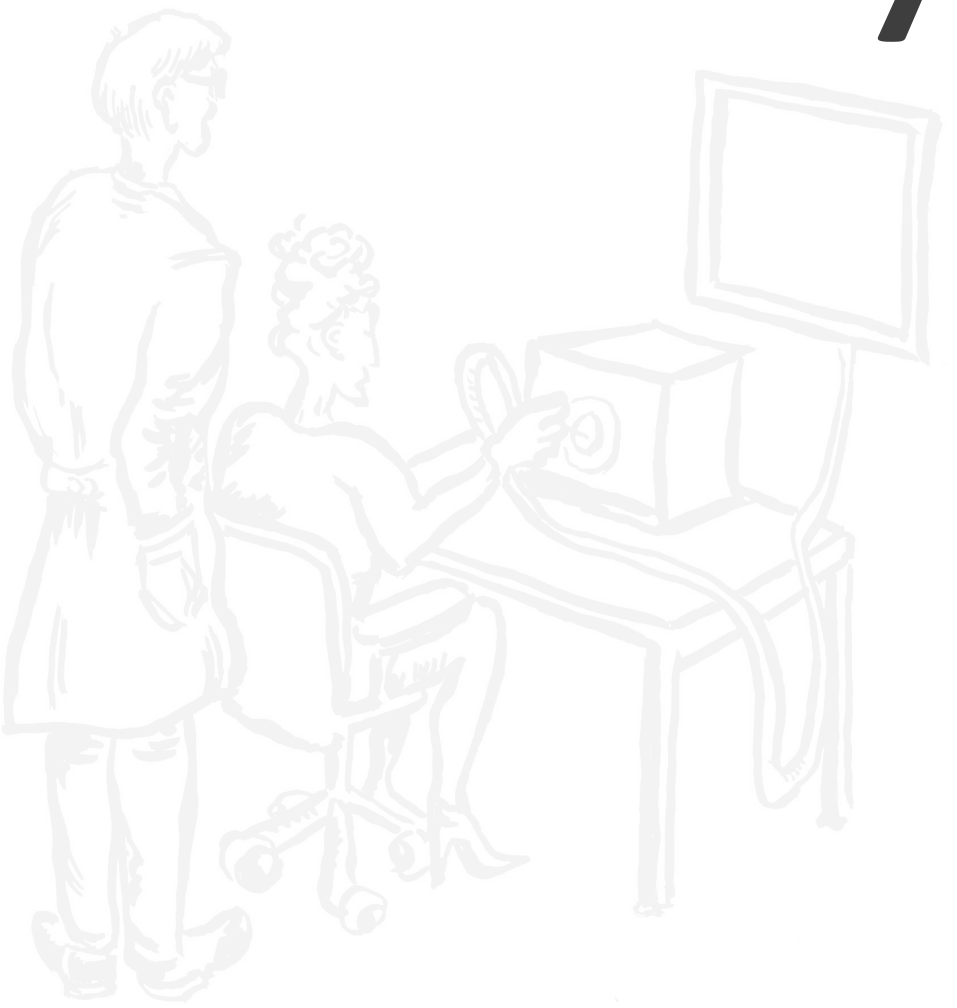
Training in Urology. Time to Test!

de Vries, A.H.

2017

document version
Publisher's PDF, also known as Version of record

7



Development and validation of the TOCO-TURBT tool: a high-stakes assessment tool that measures surgical competency in transurethral resection of bladder tumour

Anna H. de Vries, Arno M.M. Muijtjens, Hilde G.J. van Genugten, Ad J.M. Hendrikx, Evert L. Koldewijn, Barbara M.A. Schout, Cees P.M. van der Vleuten, Cordula Wagner, Irene M. Tjiam, Jeroen J.G. van Merriënboer

Submitted

ABSTRACT

Objectives

To develop and validate a high-stakes assessment tool that measures technical as well as non-technical competency in transurethral resection of bladder tumour (TURBT), as the current shift towards competency-based residency training has increased the need for objective assessment of skills.

Subjects and methods

The 'Test Objective Competency' (TOCO) - TURBT tool was designed by means of a cognitive task analysis (CTA), which included expert consensus. The tool consists of 51 items and is divided into three phases: preparatory (n=15), procedural (n=21), and completion (n=15). For validation of the TOCO-TURBT tool, two TURBT procedures were performed by 25 urologists and 51 residents in a simulated setting, and videotaped. The participants' degree of competence was assessed by a panel of 8 independent expert urologists using the TOCO-TURBT tool. Each procedure was assessed by two raters. Feasibility, acceptability and content validity were evaluated by means of a quantitative cross-sectional survey. Regression analyses were performed to assess the strength of the relation between experience and test scores (construct validity). Reliability was analyzed by generalizability-theory (G-theory).

Results

The majority of assessors and urologists indicated the TOCO-TURBT tool to be a valid assessment of competency and would support the implementation of the TOCO-TURBT assessment as a certification method for residents. Construct validity was clearly established for all outcome measures of the procedural phase (all $r > 0.5$, $p < 0.01$). G-theory analysis showed high reliability (coefficient Phi ≥ 0.8) when using the format of two assessors and two cases to assess a participant.

Conclusion

This study provides first evidence that the TOCO-TURBT tool is a feasible, valid and reliable high-stakes assessment tool for measuring surgical competency in TURBT. The tool has the potential to be used for future certification of skills for residents and urologists. The methodology of CTA might be valuable in the development of assessment tools in other areas of clinical practice.

INTRODUCTION

Traditionally, declarations of competency in surgical skills have been based on the number of cases performed and the subjective opinion of a mentor, both indicating a perception of performance rather than an actual measurement of skills. [1-3] The current shift from time-based residency training to competency-based residency training has led to a growing demand for objective assessment of skills. [4,5] Besides, several cases of technical incompetence have led to unacceptable patient morbidity and mortality, which has increased public and political pressure to evaluate surgical quality and competency. [6-8]

Objective assessment provides a basis for constructive feedback; besides, it has the potential to measure competency before granting privileges for independent clinical practice. [1,9,10] In literature, the main focus in assessment has been on the measurement of technical skills. [1,11-13] However, a surgeon should not only be competent in technical skills, as also non-technical skills (e.g. communication, leadership, decision-making, teamwork and situational awareness) play a significant role. [14,15] Up till now, few studies have focused on the concurrent assessment of technical and non-technical skills. [15,16]

A method that could be suitable for the development of an assessment tool that measures technical as well as non-technical skills is 'Cognitive Task Analysis' (CTA). [17] CTA uses a variety of interview and observation strategies that aim to capture experts' knowledge, thought processes, and decision-making during the performance of a complex task. [18-20] As experts have automated certain parts of their performance, they are no longer conscious of every step they take, which can lead to difficulties in the identification of decision points. [18,21,22] CTA offers a unique educational method to deconstruct the automated skills of experts and to identify relevant steps and decision points. [23]

Within urology, transurethral resection of bladder tumour (TURBT) is a basic but vital procedure that every urologist should master. It is the initial treatment for bladder cancer, and competency in TURBT is paramount since an accurate and radical TURBT is essential for achieving a good prognosis. [24-26] Surgical experience has been found to be predictive for recurrence after TURBT for Ta-T1 bladder cancer, and resident involvement in lower urinary tract surgery has been found to be associated with increased readmissions. [27-29] This reflects the importance of training and assessment in order to minimize potential risks for patients, and it emphasises the need for high-stakes assessment before privileges to independent clinical practice are granted. Up till now, no validated high-stakes assessment tool is available that measures technical as well as non-technical competency in TURBT.

This study aims to develop and validate a summative assessment tool for TURBT by addressing the following research questions: "What are the technical and non-technical skills necessary for a urologist to show competency in TURBT?" and "Is the newly developed Test Objective Competency (TOCO) - TURBT tool a valid and reliable assessment tool?"

SUBJECTS AND METHODS

Development of the TOCO-TURBT tool

The TOCO-TURBT tool was developed by means of a CTA between April and September 2014. The stepwise method that was followed is reflected in Figure 1.

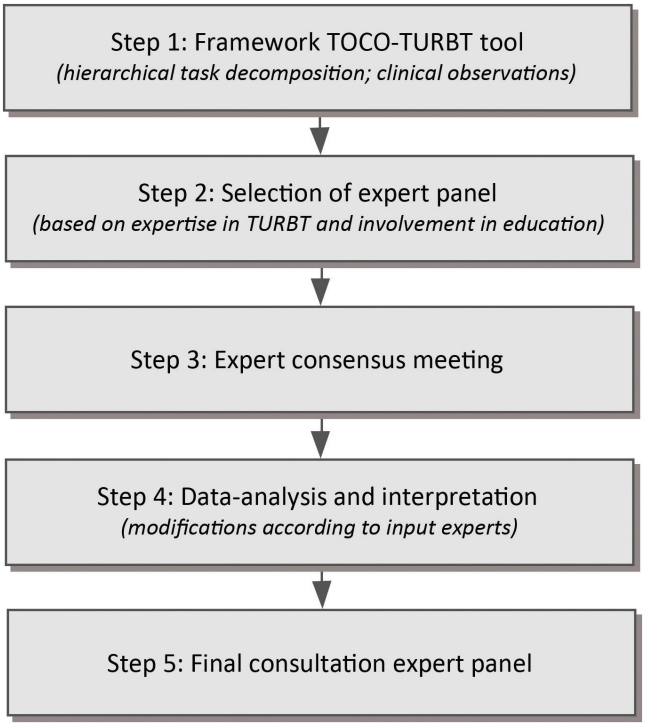


Figure 1. Stepwise method for the development of the TOCO-TURBT tool

The first step was the design of a framework that consisted of a hierarchical description of all the constituent skills that enable competent performance of TURBT (hierarchical task decomposition). This framework was designed by two urologists and two residents. Additional constituent skills were identified during 10 hours of clinical observation.

This resulted in a detailed overview of all discrete steps of the procedure, including a description of standards for acceptable performance (aiming to make the assessment less subjective), and values for rating. An expert consensus meeting was organised, in which nine expert urologists participated. Urologists were selected based on their extensive expertise in TURBT, including familiarity with international guidelines, and on their involvement in the education of urological residents. During the consensus meeting, these experts extensively discussed the content of the framework, the description of the standards for acceptable performance, and the values for rating until they reached consensus on content, relevance and completeness. Subsequently, modifications to the framework were made according to the input and comments of the experts. Finally, the adjusted version of the assessment tool was sent to the expert panel for their final consent.

The final version of the TOCO-TURBT tool consists of 51 items and is divided into three phases: preparatory (n=15), procedural phase (n=21) and completion phase (n=15) (Appendix 1). The values for rating include yes/no and a 4-point Likert scale, in which scores of three or more indicate competent performance of that particular item. [30] In addition, an assessment of overall performance (scale 1-10) and an expert global evaluation (competent / not competent) were included.

Validation of the TOCO-TURBT tool

Study design and participants

This prospective, observational and comparative study was conducted at the urology departments of seven teaching hospitals across the Netherlands. A total of 76 residents and urologists with different levels of endoscopic experience were included. All participants received a standardized verbal introduction on the use of the validated Simbla TURBT simulator [31] from one of the tutors (HvG or HdV). Subsequently, each participant performed two standardized TURBT procedures. Standardization included the use of identical bladder substrates and the resection of four bladder tumours per procedure in a predefined order. The participants were instructed to perform a complete tumour resection and to resect the tumours until 2-3mm below the surface of the bladder wall. No guidance, instruction or feedback was provided regarding the procedural steps nor regarding the technique of performance. To enable the assessment of cognitive skills (situational judgment, the participants' understanding of the procedure, salient decision points etc.), a talk-aloud protocol was created, in which participants were instructed to express all their considerations and decisions throughout the procedure out loud. [32] Performance was recorded on video, combining external views with endoscopic views. Throughout each procedure, one of the tutors (HvG or HdV) was present.

Video assessment

The video recordings were reviewed and scored with the TOCO-TURBT tool by a panel of eight independent expert urologists. To enhance the inter-rater reliability, the video assessments were preceded by an assessor meeting, in which four videos of participants with varying experience were reviewed and assessed. During this meeting, the description of standards for acceptable performance and the accompanying values for rating were clarified.

The experts were selected based on their expertise in TURBT and on their involvement in the education of urological residents. They were blinded for the participants' training status. Each procedure was assessed by a set of two raters. For practical reasons, the experts only assessed the procedural phase of the TURBT. The preparatory phase and the registration phase were assessed by two investigators (HdV and HvG), both medical doctors with a urological background. These investigators attended a similar assessor meeting prior to starting their assessments.

Assessment of feasibility, acceptability, and content validity

Data on demographics of participants, feasibility, acceptability, and content validity of the TOCO-TURBT tool were collected by means of a questionnaire, which was derived from Barton and colleagues.[8] The questionnaire was completed by the participants and the assessors. Question formats included multiple-choice questions and open-ended questions. A copy of the questionnaire can be obtained from the primary author at the reader's request.

Assessment of construct validity

The following outcome measures were used to evaluate the construct validity of the TOCO-TURBT tool: 1. Test score per phase, subdivided into PrepScore (preparatory phase), ProcScore (procedural phase), and ComplScore (completion phase). 2. Overall performance (GlobalScore, 1-10), 3. Competency score (ComptcyScore; 0 = not competent; 1 = competent), and 4. Resection time (LogTime).

Statistical analysis

The test score per phase was defined as the percentage credit points obtained over all items of that phase. Each item contributed a maximum of 1 point to the sum score. As there were scores from two assessors, the aggregate item score was obtained by calculating the mean of the two scores. For each of the two cases a proportion correct score was obtained by calculating the sum score over all the items and dividing it by the number of nonmissing items. The final proportion correct score over cases was obtained by calculating the mean of proportion correct scores of cases 1 and 2.

For the procedural phase, aggregate scores were also obtained for the outcome measure 'overall performance' and 'competency score'. Concerning the outcome measure 'resection time', a skewed distribution was anticipated. Therefore, a logarithmic transformation was applied to de-skew the distribution, and the resulting variable LogTime was used in the analysis. This variable was objectively measured by a single assessor (HdV).

Reliability

The reliability of the TOCO-TURBT tool was estimated using the generalizability theory (G-theory).[33] This approach uses a variance components analysis (G-study) to measure the contributions of all relevant factors to the total variance (variance due to variation in persons, assessors, cases, and their interactions). For the procedural phase, the design used in the G-study was assessors nested within participants-crossed-with-cases (different assessors rated each participant-case combination). Since we were interested in the absolute reliability, allowing different assessors and different cases in future assessments, the dependability index coefficient *Phi* was calculated as an indicator.[34] The formula for Phi is presented in Appendix 2a. The coefficient varies between 0 and 1, and for high-stakes assessment a *Phi* of 0.8 or higher is generally considered sufficient.[35] The variance components obtained in the G-study were used in a subsequent D-study to obtain information about the number of assessors and cases required to reach a certain level of reliability in an examination.

For the preparatory and completion phase, the reliability of the outcome measures PrepScore and ComplScore was also investigated by means of a G-theory analysis. The design of this analysis was fully crossed, as the same two assessors judged the performance of all participants in all cases. The corresponding formula for *Phi* is reflected in Appendix 2b.

Construct validity

For the assessment of construct validity, the relation between the outcome measures and the experience of participants was investigated. The experience of residents was weighted according to the levels of (in)dependent performance as recorded in their individual portfolio. For urologists, a distinction was made between partially and completely performed procedures. The equations used for the estimation of experience are described in Appendix 3.

The distribution of Experience was extremely skewed and peaked (skewness=2.8. kurtosis=8.7). Therefore, it was transformed into a logarithmic scale, and the constant 0.5 was added to prevent the argument of Log_{10} to become 0, the logarithm of 0 being undefined.

Regression analyses were performed to assess the strength of the relation between a participant's experience and the different outcome measures. The resulting regression

coefficient (b) represents the slope of the regression line, indicating the increase of the outcome measure when LogExp increases by 1. The strength of the relation is indicated by the correlation coefficient (r) of the outcome measure and LogExp. According to Cohen's classification, correlations equal to 0.1, 0.3, and 0.5 correspond to small, moderate, and large effect sizes, respectively.[36] All statistical analyses were performed using SPSS version 22.

Ethics

Ethical approval was sought from the institution's research and ethics committee. They ruled that ethical approval was not required according to the Dutch Medical Research (Human Subjects) Act, since no patients or patient details were involved. Informed consent with assurance of anonymity and confidentiality was obtained from all participants.

RESULTS

Between February and July 2015 a total of 76 participants were included in this study, 51 of whom were residents and 25 urologists. The general demographics of the participants and the assessors are described in Table 1.

Table 1. General demographics

	Age (year)	Gender (M/F)	Dexterity (right/left)	Experience in TURBT (independently performed)
PG Y0 (n=8)	26.5 (25-29)	3/5	6/2	0 (0)
PG Y1 (n=3)	28 (28)	2/1	3/0	0 (0)
PG Y2 (n=4)	29.5 (28-30)	2/2	3/1	0 (0)
PG Y3 (n=12)	30 (28-34)	8/4	12/0	0 (0-20)
PG Y4 (n=6)	30 (30-36)	1/5	5/1	0 (0-1)
PG Y5 (n=11)	32 (31-38)	8/3	9/2	6 (0-30)
PG Y6 (n=7)	34 (32-37)	2/5	7/0	6 (2-38)
Urologist (n=25)	49 (32-64)	23/3	24/1	100 (50-500)
Assessor (n=8)	54 (43-68)	8/0	8/0	350 (100-1000)

Values for variables (age and experience in TURBT) are presented as medians with range (min-max)
PG Y0 = post graduate student, not yet in training; PG Y1 – 6 = resident in training yr. 1-6

Feasibility, acceptability and content validity

All assessors and participating urologists (strongly) agreed that the TOCO-TURBT tool covered all the important aspects of the TURBT procedure, and 93% of assessors and urologists (strongly) agreed that the tool seemed to be valid. All assessors considered the assessment process to be understandable and transparent, and indicated that the assessment tool corresponds with their professional judgement regarding competence. The majority of assessors and participants (88% and 62% respectively) supported the implementation of the TOCO-TURBT assessment as a certification method before allowing residents at the end of their traineeship to independently perform a TURBT procedure on a patient. Moreover, the majority of assessors (88%) and participating urologists (95%) agreed that the TOCO-TURBT tool could be used for the certification of urologists.

Reliability

The estimated variance components for the five outcome measures obtained in the G-analyses are shown and explained in Appendix 4. For all measures, the largest component appeared to be the person variance (the variance of interest). This indicates that the person variance has the largest influence on the outcome measures, which is favourable for the reliability.

Table 2 shows the absolute reliability (coefficient *Phi*) resulting from the variance components reflected in Appendix 4. As the table shows, the reliability is improved by an increase in the number of assessors and/or cases. In the current set-up of the TOCO-TURBT tool, two cases and two assessors were used for each assessment. With this set-up, satisfactory levels of reliability were obtained for all five measures (range 0.79-0.87).

Table 2. Reliability (coefficient Phi) of the outcome measures in relation to the numbers of assessors (Na) and cases (Nc) used to assess a participant with the TOCO-TURBT tool.

	Preparatory phase			Procedural phase									Completion phase		
	PrepScore			ProcScore			GlobalScore			ComptcyScore			ComplScore		
Na	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Nc															
1	.74	.79	.81	.61	.73	.78	.65	.77	.83	.51	.65	.71	.62	.66	.67
2	.83	**.87**	.89	.76	**.84**	.87	.79	**.87**	.91	.68	**.79**	.83	.76	**.79**	.80
3	.86	.90	.91	.83	.89	.91	.85	.91	.93	.76	.85	.88	.82	.85	.86

For high-stakes assessment a *Phi* of 0.8 or higher is generally considered sufficient.[35]

Construct validity

Table 3 shows the descriptives and the results of the regression analyses for the six outcome measures, with LogExp as independent variable. For one participant time registrations for both cases were lacking, hence, for LogTime only 75 observations were available.

The regression results for the preparation and completion phase (right panel of Table 3) showed no significant statistical relation between the participant's experience and the concurrent PrepScore and ComplScore. For the Procedural phase, all outcome measures show a statistically significant linear relation with experience (LogExp). Moreover, the values of correlation (r) are in the range of 0.61-0.72, indicating large effect sizes. For LogTime, a negative relation with experience was found (b=-0.076), which is in agreement with the expectation that resection time in general will be lower for more experienced participants. The relation of this measure is not as strong as that of the other three procedural phase measures, but still substantial with a correlation (r) equal to -0.34, which indicates an effect of moderate size. The results for the procedural phase outcome measures are very supportive of the construct validity of the TOCO-TURBT tool.

Table 3. Descriptives and regression analysis results for the outcome measures with the logarithm of experience as independent variable.

Phase	Performance measure	Descriptives[1)]					Regression with LogExp[2)]				
		M	SD	min	max	N	b	r	p	95% CI lo	95% CI hi
Preparatory	PrepScore (0-100)	63	11	27	88	76	2.3	.17	.148	-0.8	5.4
Procedural	ProcScore (0-100)	65	17	10	91	76	12	.61	.001	9	16
	GlobalScore (1-10)	6.3	1.7	2.0	8.5	76	1.5	.72	.001	1.2	1.8
	ComptcyScore (0-1)	.56	.41	.00	1.00	76	.35	.72	.001	.27	.43
	LogTime (min)	1.91	.19	1.30	2.38	75	-.076	-.34	.003	-.13	-.03
Completion	ComplScore (0-100)	60	12	31	88	76	1.4	.10	.395	-1.8	4.6

[1)] M: mean, SD: standard deviation, min/max: minimum/maximum, N: number of subjects with non-missing value.

[2)] b: regression coefficient, r: correlation, p: p-value, lo-hi: lower and upper boundary of the 95% confidence interval of b.

DISCUSSION

This study presents the development and validation of the TOCO-TURBT tool, a high-stakes assessment tool for TURBT that measures technical as well as non-technical competency. This study clearly established the feasibility, content validity, and construct validity as well as the reliability of the TOCO-TURBT tool. This indicates that the TOCO-TURBT tool has the potential to be used for high-stakes assessment, such as certification of residents and relicensing of urologists.

The TOCO-TURBT tool was developed using the methodology of CTA, a constructive and systematic approach that has enabled the identification and definition of all the relevant technical and non-technical aspects of the TURBT procedure. The additional value of using the CTA methodology is that it captures automated expert knowledge that would otherwise have been lost. [37] Although this method is time-consuming and labour-intensive, it is recognized that a constructive approach to instrument development (including e.g. an a priori conceptual framework and use of expert consensus) helps to ensure that it captures the concept of interest, thereby increasing its content validity and generalizability. [38]

Before a new assessment tool can be implemented in clinical practice, its feasibility, acceptability, validity, and reliability must be established. [39] Feasibility and acceptability were confirmed, as the majority of assessors and participants indicated that they would support the implementation of the TOCO-TURBT assessment as a certification method for residents. For the assessment of construct validity (the capability of the assessment tool to differentiate between different levels of experience), the broad classification of 'novices', 'intermediates', and 'experts' is generally used. This classification seems somewhat arbitrary and lacks precision. Therefore we used a different approach, derived from the validation study of the Program for Laparoscopic Urological Skills, performed by Tjiam and colleagues. [40] Instead of using 'experience' as a categorical variable, it was used as a continuous variable, expressed as the absolute number of TURBT procedures performed. In order to optimize the accuracy of the estimation, the experience of residents was weighted according to the levels of supervised or independent performance as recorded in their individual portfolios.

Construct validity was clearly demonstrated for the procedural phase of the TOCO-TURBT tool, but it was lacking for the preparatory and completion phase. The absence of construct validity in the preparatory and completion phase did not come as a surprise, and can be explained by the fact that the majority of items included in these phases are 'checkpoints', for which little experience or technical skill is needed. These two phases can be seen as the 'basics' of a TURBT procedure. The distinction between a novice and an expert becomes apparent in the procedural phase, where the actual TURBT is assessed, including the resection skills, handling complications etc.

The procedural phase of the TURBT procedures was assessed by a panel of expert urologists, whereas the other two phases were assessed by two general doctors. Ideally, the experts would have assessed the complete procedure, but unfortunately this was not possible due to time constraints. Still, we consider the current assessment approach to be valid, as the items of the preparation and completion phase do not need an 'expert eye' per se to be adequately assessed, whereas this is required for the procedural phase.

The reliability of the TOCO-TURBT tool was evaluated using Generalizability theory. The benefit of this method is that it measures the contributions of all relevant factors to the final result, thereby allowing an estimation of the size of the relevant variables that influence the reliability of a given score.[41] The results of our study showed that the TOCO-TURBT tool is a reliable assessment tool, with two assessors and two trials being sufficient to reach substantial reliability.

To our knowledge, the TOCO-TURBT tool is the first assessment tool in the field of Urology that has the potential to be used for high-stakes assessment in the future. Besides the potential use of the TOCO-TURBT tool for certification and relicensing, it could also be used in the light of the concept "Entrustable Professional Activities" (EPAs).[42] An EPA is an activity residents can be trusted to perform competently in different stages of training. This concept translates competencies into clinical practice and enables supervisors to determine when a resident can be trusted to perform specific procedures with minimal supervision or without supervision. The TOCO-TURBT tool could be used to objectify a resident's performance throughout the different stages of independence. For this, the components of the TURBT procedure that residents should master during different stages of their residency should be determined first.

The validation of the TOCO-TURBT tool was conducted in a simulation setting. The advantage of this approach is that the research setting was completely standardized, enabling a true comparison of performance without any interfering confounding factors. A drawback of this approach is that certain constituent skills, such as performing haemostasis, could not be simulated on the simulator and were only assessed in a cognitive way. Therefore, extrapolation of the use of this assessment tool to a clinical setting is desirable, especially for the purpose of certification or relicensing. This will be the topic of a future study, as will be the determination of pass/fail standards.

CONCLUSION

This study provides first evidence that the TOCO-TURBT tool is a feasible, valid and reliable high-stakes assessment tool for measuring technical and non-technical competency in TURBT. It has the potential to be used for the certification of skills of residents and urologists. The method of CTA might be of value in the development of assessment tools in other areas of clinical practice.

ACKNOWLEDGMENTS

The authors gratefully acknowledge all the experts who participated in the development and validation of the TOCO-TURBT tool, the residents and urologists who were willing to participate in this study, Ron Hoogeboom for statistical support, and Lisette van Hulst for editorial assistance.

REFERENCES

(1) Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Ann Surg 2008 Feb;247(2):372-379.

(2) Zevin B, Bonrath EM, Aggarwal R, Dedy NJ, Ahmed N, Grantcharov TP, et al. Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. J Am Coll Surg 2013 May;216(5):955-965.e8; quiz 1029-31, 1033.

(3) Sarker SK, Maciocco M, Zaman A, Kumar I. Operative performance in laparoscopic cholecystectomy using the Procedural-Based Assessment tool. Am J Surg 2010 Sep;200(3):334-340.

(4) Hampton T. Efforts seek to develop systematic ways to objectively assess surgeons' skills. JAMA 2015 Feb 24;313(8):782-784.

(5) de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. Br J Surg 2016 Feb;103(3):300-306.

(6) Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ 2004 May 25;170(11):1678-1686.

(7) Vincent C, Moorthy K, Sarker SK, Chang A, Darzi AW. Systems approaches to surgical quality and safety: from concept to measurement. Ann Surg 2004 Apr;239(4):475-482.

(8) Barton JR, Corbett S, van der Vleuten CP, English Bowel Cancer Screening Programme, UK Joint Advisory Group for Gastrointestinal Endoscopy. The validity and reliability of a Direct Observation of Procedural Skills assessment tool: assessing colonoscopic skills of senior endoscopists. Gastrointest Endosc 2012 Mar;75(3):591-597.

(9) Miskovic D, Ni M, Wyles SM, Kennedy RH, Francis NK, Parvaiz A, et al. Is competency assessment at the specialist level achievable? A study for the national training programme in laparoscopic colorectal surgery in England. Ann Surg 2013 Mar;257(3):476-482.

(10) Palter VN, MacRae HM, Grantcharov TP. Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. Am J Surg 2011 Feb;201(2):251-259.

(11) Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery 2005 Oct;138(4):640-7; discussion 647-9.

(12) Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997 Feb;84(2):273-278.

(13) Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005 Jul;190(1):107-113.

(14) Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. Surgery 2006 Feb;139(2):140-149.

(15) Hull L, Arora S, Aggarwal R, Darzi A, Vincent C, Sevdalis N. The impact of nontechnical skills on technical performance in surgery:

a systematic review. J Am Coll Surg 2012 Feb;214(2):214-230.

(16) Shepherd W, Arora KS, Abboudi H, Shamim Khan M, Dasgupta P, Ahmed K. A review of the available urology skills training curricula and their validation. J Surg Educ 2014 May-Jun;71(3):289-296.

(17) van Merrienboer JJG, Kirschner PA. Ten steps to complex learning. A systematic approach to four-component instructional design. New York/London: Routledge 2013(Second edition):78-111.

(18) Clark R, Feldon D, van Merrienboer J, Yates K. Cognitive task analysis

 . Handbook of research on educational communications and technology: Taylor and Francis Group; 2008. p. 577-593.

(19) Yates K, Feldon D. Advancing the practice of cognitive task analysis: a call for taxonomic research. Theoretical Issues in Ergonomics Science, 2011(11):1464-1536.

(20) Salas E, Cannon-Bowers JA. The science of training: a decade of progress. Annu Rev Psychol 2001;52:471-499.

(21) Clark R, Estes F. Cognitive task analysis for training. Int J Educ Res 1996;25:403-4017.

(22) Gagne E, Yekovich C, Yekovich FR. Cognitive Psychology of School Learning. New York: Harper Collins; 1993.

(23) Sullivan ME, Ortega A, Wasserberg N, Kaufman H, Nyquist J, Clark R. Assessing the teaching of procedural skills: can cognitive task analysis add to our traditional teaching methods? Am J Surg 2008 Jan;195(1):20-23.

(24) Babjuk M, Burger M, Zigeuner R, Shariat SF, van Rhijn BW, Comperat E, et al. EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2013. Eur Urol 2013 Oct;64(4):639-653.

(25) Schout B, Dolmans V, Bemelmans B, Schoot D, Scherpbier A, Hendrikx A. Teaching diagnostic and therapeutic procedures of bladder pathology using a newly developed pig bladder model. J Endourol 2008 Nov;22(11):2547-2553.

(26) Brausi M, Collette L, Kurth K, van der Meijden AP, Oosterlinck W, Witjes JA, et al. Variability in the recurrence rate at first follow-up cystoscopy after TUR in stage Ta T1 transitional cell carcinoma of the bladder: a combined analysis of seven EORTC studies. Eur Urol 2002 May;41(5):523-531.

(27) Allard CB, Meyer CP, Gandaglia G, Chang SL, Chun FK, Gelpi-Hammerschmidt F, et al. The Effect of Resident Involvement on Perioperative Outcomes in Transurethral Urologic Surgeries. J Surg Educ 2015 May 20.

(28) Jancke G, Rosell J, Jahnson S. Impact of surgical experience on recurrence and progression after transurethral resection of bladder tumour in non-muscle-invasive bladder cancer. Scand J Urol 2014 Jun;48(3):276-283.

(29) Roupret M, Yates DR, Varinot J, Phe V, Chartier-Kastler E, Bitker MO, et al. The presence of detrusor muscle in the pathological specimen after transurethral resection of primary pT1 bladder tumors and its relationship to operator experience. Can J Urol 2012 Oct;19(5):6459-6464.

(30) Barton JR, Corbett S, van der Vleuten CP, English Bowel Cancer Screening Programme, UK Joint Advisory Group for Gastrointestinal Endoscopy. The validity and reliability of a Direct Observation of Procedural Skills assessment tool: assessing colonoscopic skills of senior endoscopists. Gastrointest Endosc 2012 Mar;75(3):591-597.

(31) de Vries AH, van Genugten HG, Hendrikx AJ, Koldewijn EL, Schout BM, Tjiam IM, et al. The Simbla TURBT Simulator in Urological Residency Training: From Needs Analysis to Validation. J Endourol 2016 Jan 22.

(32) Luker KR, Sullivan ME, Peyre SE, Sherman R, Grunwald T. The use of a cognitive task analysis-based multimedia program to teach surgical decision making in flexor tendon repair. Am J Surg 2008 Jan;195(1):11-15.

(33) Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Med Educ 2002 Oct;36(10):972-978.

(34) Brennan RL. Generalizability Theory. New York: Springer Verlag; 2001.

(35) Fraenkel JR, Wallen NE.
How to design and evaluate research in education. 7th ed. Boston: McGraw Hill; 2009.

(36) Cohen J.
Statistical power analysis for the behavioural sciences.
. London: Lawrence Erlbaum; 1988.

(37) Yates K, Sullivan M, Clark R. Integrated studies on the use of cognitive task analysis to capture surgical expertise for central venous catheter placement and open cricothyrotomy. Am J Surg 2012 Jan;203(1):76-80.

(38) Walsh CM, Ling SC, Khanna N, Cooper MA, Grover SC, May G, et al. Gastrointestinal Endoscopy Competency Assessment Tool: development of a procedure-specific assessment tool for colonoscopy. Gastrointest Endosc 2014 May;79(5):798-807.e5.

(39) Wragg A, Wade W, Fuller G, Cowan G, Mills P. Assessing the performance of specialist registrars. Clin Med 2003 Mar-Apr;3(2):131-134.

(40) Tjiam IM, Persoon MC, Hendrikx AJ, Muijtjens AM, Witjes JA, Scherpbier AJ. Program for laparoscopic urologic skills: a newly developed and validated educational program. Urology 2012 Apr;79(4):815-820.

(41) Tjiam IM, Schout BM, Hendrikx AJ, Muijtjens AM, Scherpbier AJ, Witjes JA, et al. Program for laparoscopic urological skills assessment: setting certification standards for residents. Minim Invasive Ther Allied Technol 2013 Feb;22(1):26-32.

(42) Ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013 Mar;5(1):157-158.

Appendix 1.

Summative TOP – TURBT Assessment Form

Test Objective Performance – Transurethral Resection Bladder Tumor

*** Scale and criteria key**

4 Good	Highly skilled performance
3 Acceptable	Performance of acceptable standard but could be improved
2 Marginal	Some major standards not yet met, aspects to be improved
1 Poor	Accepted standard not yet met, frequent errors uncorrected
Y / N	Performed / Not performed
N/A	Not Applicable

Code candidate: ……..
Assessor: ……
Signature: ……..
Date (DD/MM/YYYY): ……..

Criteria Preparatory phase	Score*	Comments
Patient related details		
Checks correct patient (name and date of birth)	Y / N	
Checks indication	Y / N	
Checks allergies	Y / N	
Checks urinary culture / antibiotic prophylaixs	Y / N	
Checks anticoagulation and handles accordingly	Y / N	
Checks indication n. obturatorius block and handles accordingly	Y / N	
Material related details (non-sterile)		
Checks equipment (up to date?)	Y / N	
Checks if all instruments are present and work properly	Y / N	
Checks/adjusts settings diathermy	Y / N	
Checks if appropriate irrigation fluid is present	Y / N	
Patient preparation		
Positions/checks positioning of the patient	Y / N	
Performs bimanual examination before start procedure	Y / N	
Applies antiseptic solution and drapes the patient/checks if this is adequately performed	Y / N	
Continuous variables		
Has an adequate leading role during time out procedure	Y / N	
Communicates clearly and pleasant with patient and operating team	Y / N	

Criteria procedural phase	Score*	Comments
Material related details (sterile)		
Assembles the instruments and connects the tubes correctly and competently	1| 2| 3| 4	
Adjusts the light settings (1pt), focuses the camera (2pt) and performs white balance (1pt) before the start of procedure	1| 2| 3| 4	
Introduction and inspection bladder		
Instilles lubricant into the meatus and (optional) on the final end of the obturator	Y / N	
Introduces the cystoscope under sight smoothly and safely, while inspecting the complete urethra	Y / N	
Introduces the obturator blind, smoothly and safely	Y / N | N/A	
Inspects the complete bladder using a systematic approach	1| 2| 3| 4	
Localizes the tumor(s) that have to be resected	1| 2| 3| 4	
Identifies the left and right ureteral orifice	Y / N	
Tumor resection	**Score***	**Comments**
Resection tumor(s)		
Applies an adequate strategy in tumor approach (prioritising and systematics)	1| 2| 3| 4	
Resects the tumor tissue from cranial to caudal	Y / N	
Resects strokes of adequate depth and length	1| 2| 3| 4	
Has adequate speed / progression of tumor resection	1| 2| 3| 4	
Prevents complications adequately	1| 2| 3| 4	
Handles complications adequately	1| 2| 3| 4	
Judges end result accurately, including inspection of the complete bladder	Y / N	
Continuous variables		
Smoothly changes instruments during procedure	Y / N	
Continuously maintains orientation in the bladder	1| 2| 3| 4	
Performs smooth and accurate hemostasis during complete procedure	1| 2| 3| 4	
Applies adequate technique of evacuating tissue chips	Y / N	
Pays close attention to ergonomics	Y / N	
Reacts adequate on information/messages from the complete operating team and communicates clearly	Y / N	

Criteria Completion phase	Score*	Comments
Completion		
Inserts transurethral catheter adequately	Y / N	
Checks drainage and connects continuous flow system if indicated	Y / N	
Performs a debriefing in presence of the complete operating team including		
Check count of operating materials with assisting nurse	Y / N	
Check tissue collection	Y / N	
Discussion of surgical complications	Y / N	
Discussion/organization postoperative policy	chemo Y/N; AB Y/N; anticoag Y/N; CAD policy Y/N; pain policy Y/N	
Communication		
Informs patient and/or family after procedure	Y / N	
Registration		
Completes operating report, including details of		
Description of all procedural steps	positioning Y/N; type anesth Y/N; BMT Y/N; UCS incl descr. location tumor Y/N; orifices Y/N resection Y/N; hemost Y/N; remov and coll tissue chips Y/N; CAD Y/N; cc balloon Y/N	
Radicalty of tumor resection	Y / N	
Complications	Y / N	
Completes pathology form in detail, including		
Type of material	Y / N	
Number of pots including description per pot	Y / N	
Research question	malignancy Y/N; type tumor Y/N; grade Y/N; invasion Y/N; presence detrusor Y/N	
Postoperative registration		
Registers postoperative policy in patient file, including details of:	complications Y/N; chemo Y/N; CAD policy Y/N; AB Y/N; pain policy Y/N; anticoag Y/N	
Completes the financial registration of the procedure	Y / N	

Overall evaluation of performance *
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

** Scale 1-10 to be interpreted as school grade*

Expert global evaluation *
☐ This candidate **should** be certified for level 2 TURBT
☐ This candidate **should not** be certified for level 2 TURBT
☐ This candidate **should** be certified for level 4 TURBT
☐ This candidate **should not** be certified for level 4 TURBT

** Level 2: residents at the beginning of their first year of urological residency training (exposure: <5 TURBT assisted)*

** Level 4: residents at the end of their four years of urological residency training*

Appendix 2a. Formula for *Phi* used in the G-analysis of the procedural phase

$$Phi = \frac{V_p}{V_p + V_c / N_c + V_{pc} / N_c + V_{a:pc} / (N_a \times N_c)}$$

Appendix 2b. Formula for *Phi* used in the G-analysis of the preparatory and completion phase

$$Phi = \frac{V_p}{V_p + (V_c + V_{pc}) / N_c + (V_a + V_{pa}) / N_a + (V_{ca} + V_{pca}) / (N_c \times N_a)}$$

Appendix 3a. Equation for the estimation of residents' experience in TURBT

Experience=0,25xNassist+0,50xNsup1+0,75xNsup2+1.0xNind

Nassist = no. of TURBT procedures performed as an *assistant*
Nsup1 = no. of TURBT procedures performed s a *surgeon under close supervision*
Nsup2 = no. of TURBT procedures performed as a *surgeon under limited supervision*
Nind = no. of TURBT procedures performed *independently*

Appendix 3b. Equation for the estimation of urologists' experience in TURBT

Experience=0,5xNpart+1.0xNcompl

Npart: no. of TURBT procedures *partially* performed
Ncompl: no. of TURBT procedures *completely* performed

Appendix 4. Variance components of the TOCO-TURBT tool outcome measures

Performance measure	Preparatory phase and Completion phase						
	Vp	Vc	Va	Vpc	Vpa	Vca	Vpca
PrepScore (0-100)	113.8	0.0	4.9	20.1	2.5	0.6	12.2
ComplScore (0-100)	110.3	5.1	0.4	42.8	1.0	0.0	18.4

	Procedural phase			
	Vp	Vc	Vpc	Va:pc
ProcScore (1-100)	224.7	4.5	22.0	115.6
GlobalScore (1-10)	2.501	0.109	0.000	1.243
ComptcyScore (0-1)	0.128	0.000	0.016	0.105

Variance component	***Description***
Vp	Systematic variation among participants (residents and urologists)
Vc	Systematic variation among cases
Va	Systematic variation among assessors
Vpc	Variation of difficulty of cases and variability of participants across cases
Vpa	Variation of assessors across different participants
Vpca	All remaining variability (unexplained error)

Explanation

- The person variance V_p was the variance of interest, the other components represented error variances.
- The components case variance (V_c) and assessor varianc (V_a) were relatively small, which indicated that the two cases were of comparable difficulty, and the assessors showed comparable levels of stringency.
- For Complscore, the person-case variance (V_{pc}) was of considerable size. This indicated a substantial case-specificity: while the difficulty of the cases was found to be comparable, for the same person the Complscore on the two cases may differ considerably.
- For ComptcyScore, the assessor within person-case component ($V_{a:pc}$) was almost equal to the person variance V_p. This indicated that for this measure the assessor specificity was substantial. However, when the exam consists of two cases, each assessed by two raters, the contribution of variance $V_{a:pc}$ to the error variance was reduced by a factor $N_{assessors} \times N_{cases} = 4$.
- For ProcScore and GlobalScore the assessor-specificity was also considerable with $V_{a:pc}$ approximately half as large as V_p, but similarly reduction of the error variance can be obtained by increasing the number of cases and raters.